Improving Risk Grouping Rules for Prostate Cancer Patients with Optimization

L.Churilov[1], A.M.Bagirov[2], D.Schwartz[1,3], K.Smith[1], and M.Dally[4]

[1]*School of Business Systems, Monash University, Victoria 3800, Australia*
[2]*Centre for Informatics and Applied Optimization, School of Information Technology and Mathematical Sciences, The University of Ballarat, Victoria 3353, Australia*
[3]*Department of Industrial Engineering, Faculty of Physical Sciences and Mathematics, University of Chile, Republica 701, Santiago, Chile*
[4]*William Buckland Radiotherapy Department, The Alfred, Victoria 3000, Australia*
E-mail：*Leonid.Churilov@infotech.monash.edu.au*

Abstract

Data mining techniques provide a popular and powerful toolset to address both clinical and management issues in the area of health care. This paper describes the study of assigning prostate cancer patients into homogenous groups with the aim to support future clinical treatment decisions. The cluster analysis based model is suggested and an application of non-smooth non-convex optimization techniques to solve this model is discussed. It is demonstrated that using the optimization based approach to data mining of a prostate cancer patients database can lead to generation of a significant amount of new knowledge that can be effectively utilized to enhance clinical decision making.

1. Introduction

Use of data mining techniques has become widespread in the area of medical and health care. Problems of both health management and of clinical nature are now the focus of successful research efforts by many data mining researchers. Various data mining techniques are used in the area of clinical decision support, and, in particular, cancer diagnostics and prognostics.

Mangasarian et al [10] discuss the applications of linear programming to the problem of clinical classification of patients with breast cancer. Bellazi et al [5] present the use of data mining tools to derive a prognostic model of the outcome of resectable hepatocellular carcinoma. Land et al [8] discuss a new neural network technology developed to improve the diagnosis of breast cancer using mammogram findings, while Walter and Mohan [12] describe *ClaDia*, a learning classifier system applied to the Wisconsin breast cancer data set, using a fuzzy representation of the rules, a median based fuzzy combination rule, and separate subpopulations for each class. Setiono [11] presents an algorithm that extracts classification rules from trained neural networks and discusses its application to breast cancer diagnosis, as well as describing how the accuracy of the networks and the accuracy of the rules extracted from them can be improved by a simple pre-processing of the data.

As far as prostate cancer is concerned, Zupan et al [14] propose a schema that enables the use of classification methods, including machine learning classifiers, for survival analysis of prostate cancer patients, while Zhang and Zhang [13] develop and validate *ProstAsure* - a neural network-derived algorithm which analyzes the profile of multiple serum tumor markers and produces a single-valued diagnostic index for early detection of prostate cancer.

Most of the different approaches to the problem of clustering analysis suggested in the literature are mainly based on statistical, neural network, and machine learning techniques. An excellent survey of existing approaches is provided in Jain et al [6]. At the same time, where are relatively few attempts to utilize optimization techniques for this purposes (Mangasarian [9]).

Bagirov et al [4] propose the global optimization approach to clustering and demonstrate how the supervised data classification problem can be solved via clustering. The objective function in this problem is both non-smooth and non-convex and this function has a large number of local minimizers. Problems of this type are quite challenging for general-purpose global optimization techniques. Due to a large number of variables and the complexity of the objective function, general-purpose global optimization techniques, as a rule, fail to solve such problems. It is very important, therefore, to develop optimization algorithms that allow the decision maker to find "deep" local minimizers of the objective function. Such "deep" local minimizers provide a good enough description of the dataset under consideration as far as clustering is concerned. The optimization algorithm discussed in this paper belongs to this type and is based on non-smooth optimization techniques.

The *objective* of this paper is to demonstrate how optimization-based clustering techniques discussed in Bagirov and Churilov [3] can be utilized to cluster the prostate cancer patients into risk-homogeneous patient groups in order to support future treatment decisions.

Patient's age, tumor stage, pathology Gleason score, and PSA (prostate-specific antigen) levels in blood are used to generate clusters that reveal interesting differences in patients' future health and survival. A detailed case study is presented to demonstrate the quality of knowledge generated by this process. It is suggested that the proposed approach can, therefore, be seen as an evidence-based predictive tool with significant knowledge generation capabilities.

This paper is organized as follows: section 2 describes the case study settings and provides the elementary analysis of the case data; in section 3 the optimization approach to clustering the prostate cancer database is described and the corresponding algorithm is presented; the results of the study are discussed in section 4; while section 5 provides the summary and conclusions for this study.

2. Problem Context and Data Analysis

The William Buckland Radiotherapy Center (WBRC) in metropolitan Melbourne, Australia had been selecting the prostate cancer patients for various treatment approaches using risk groupings derived from a review of published predictive models as shown in Table 1.

Patients' diagnostic statistics used in this case study includes Gleason Score (1-best, 10-worst), tumor stage (1a-best, 4-worst), and the PSA level at the time of diagnosis, as well as patient's age. The database consists of 258 de-identified records of hormone-naïve patients who underwent external beam radiotherapy at some stage between 1/1/1990 and 31/12/1997. As part of clinical and PSA follow-up that happened at 3,6,9,12,18,and 24 months, then annually, the biochemical (PSA) failure-free survival (bFFS) was calculated from the beginning of radiotherapy until the biochemical failure. Biochemical (PSA) failure was defined as per ASTRO [1] consensus statement or any event that prompted androgen deprivation

therapy. The concept of risk in this situation is directly linked to the chance of a bFFS over a given period of time (usually, 5 years) – i.e. the higher the chance of biochemical (PSA) failure-free survival for a given patient, the lower is the corresponding clinical risk, and the less aggressive is the clinical treatment strategy that can be implemented.

Applying the rules shown in Table 1 to the initial dataset, it becomes apparent that according to WBRC classification 8.2% of all the patients belong to the "low risk" group, 51.6% belong to the "intermediate risk" group, while the remaining 38.3% of the patients belong to the "high risk" group. These findings are summarized in Table 2. Note that this classification presents a relatively high percentage of patients in the "intermediate" group, thus complicating the decisions regarding appropriate treatment for many patients.

It is important to note that, as indicated in Table 1, the classification based on three aggregate groups is derived by applying seven rules. If all the seven rules are allowed to generate their corresponding risk groups, the new risk classification table can be obtained (Table 3).

Note that separating the data into seven groups according to the rules used by WBRC reveals one very important inconsistency specifically demonstrating that the bFFS rate for the group High2 is, in fact, higher than the bFFS rate for the group Intermediate3. This directly contradicts the way the concept of clinical risk is intuitively defined and raises the question of the appropriateness of the set of rules used by the WBRC to identify specific risk groups. Note also that this classification, although stipulating better "granularity" of risk groups, is fundamentally based on the set of (sometimes inappropriate) rules reported in the literature rather than "data-driven" rules that are elicited from the practical history observed in WBRC.

The latter can provide rich material for a data mining or, more specifically, clustering study. It is important to note that the natural desire of a clinical decision maker is to ensure that the following two conditions are met as closely as practical:

1. The bFFS measure for "high risk" group is as low as possible, while for "low risk" group is as high as possible
2. The size of the "intermediate risk" group is minimal thus making sure that the patient is either a "high risk" or a "low risk" one in order to make the future treatment decision as tightly linked to patients risk attribute as possible

These two conditions provide the decision maker with two objectives that are not necessarily mutually attainable and can even be conflicting. In such cases there will be a need to analyze the value trade-off depending on the attitudes of a given clinician or groups of clinicians (Keeney [7]). Such an analysis will have to include the definition of clinically acceptable numerical values of bFSS for "high", "intermediate", and "low" risk groups. Also, the goodness of a given data-mining (clustering) tool in these settings can be determined by how closely the two objectives are simultaneously met by the risk grouping produced by the application of a given tool.

3. The non-smooth optimization approach to solving clustering problems

The subject of cluster analysis is the unsupervised classification of data and discovery of relationships within the data set without any guidance. The basic principle of identifying these hidden relationships is that if input patterns are similar, they should be grouped together. Two inputs are regarded as similar if the distance between these two inputs (in multidimensional input space) is small.

With this notion in mind, consider set A that consists of r n-dimensional vectors $a^i=(a^i_1,\ldots, a^i_n)$, $i=1,\ldots,r$, where n is the number of data fields/attributes. The aim of clustering is to represent this set as the union of q clusters. Since each cluster can be described by the location of its center, it is

instrumental to locate a cluster's center in order to adequately describe the cluster itself. Thus, we address the problem of finding q points that serve as centers of corresponding clusters.

Consider now an arbitrary set X, consisting of q points $x^1,\ldots,x^q$. The distance $d(a^i,X)$ from a point a^i that belongs to the set A to the set X is defined by

$$d(a^i, X) = \min_{s=1,\ldots,q} \left\| x^s - a^i \right\| \quad \textbf{(1)}$$

where

$$\|x\|_p = \left(\sum_{l=1}^{n} |x_l|^p \right)^{1/p}, \quad 1 \le p < \infty, \quad \textbf{(2)}$$
$$\|x\|_\infty = \max_{l=1,\ldots,n} |x_l|$$

The deviation $d(a^i,X)$ from the set A to the set X can be calculated using the formula

$$d(A, X) = \sum_{i=1}^{r} d(a^i, X)$$
$$= \sum_{i=1}^{r} \min_{s=1,\ldots,q} \left\| x^s - a^i \right\| \quad \textbf{(3)}$$

Thus, as far as optimization approach is concerned, the cluster analysis problem can be reduced to the following mathematical programming problem:

$$\min f(x^1,\ldots,x^q)$$
$$s.t.\ (x^1,\ldots,x^q) \in \Re^{n \times q} \quad \textbf{(4)}$$

where

$$f(x^1,\ldots,x^q) = \sum_{i=1}^{r} \min_{s=1,\ldots,q} \left\| x^s - a^i \right\| \quad \textbf{(5)}$$

If $q>1$, the objective function f in the problem (4) is non-convex and non-smooth. Note that the number of variables in the optimization problem (4) is $q \times n$. If the number q of clusters and the number n of data attributes are large, the decision maker is facing a large-scale global optimization problem. Moreover, the form of the objective function in this problem is complex enough not to become amenable to the direct application of general-purpose global optimization methods. Therefore, in order to ensure the practicality of the optimization approach to clustering, the proper identification and use of local optimization methods with the special choice of a starting point is very important. Clearly, such an approach does not guarantee the globally optimal solution the problem (4). On the other hand, this approach allows one to find a "deep" minimum of the objective function that, in turn, provides a good enough clustering description of the dataset under consideration.

Note also that the meaningful choice of the number of clusters is very important for clustering analysis. It is difficult to define a priori how many clusters represent the set A under consideration. In order to increase the knowledge generating capacity of the resulting clusters, the optimization based approach discussed in this paper adopts the following strategy: starting from a small enough number of clusters q, the decision maker has to gradually increase the number of clusters for the analysis until certain termination criteria motivated by the underlying decision making situation is satisfied.

As far as optimization is concerned, this means that if the solution of the corresponding optimization problem (4) is not satisfactory, the number of clusters q should be iteratively increased so that the problem (4) should be solved with $q+1$ clusters, etc, until some termination criterion is met. This implies that one needs to solve repeatedly arising global optimization problems (4) with different values of q - the task even more challenging than solving a single global optimization problem. In order to avoid this difficulty, a step-by-step calculation of clusters is implemented in the optimization algorithm discussed below.

It is also important to note that the form of the objective function in the problem (4) allows one to significantly reduce both the

number of attributes (feature selection) and the number of records in a dataset. Although not essential in the case reported in this paper, this ability proves very important for mining large datasets. The way the proposed algorithm utilizes these features is discussed in detail in Bagirov et al [4] and Bagirov and Churilov [3].

The following optimization algorithm originally presented in Bagirov and Churilov [3] is used to cluster the prostate cancer patients for the purposes of this study:

Optimization Algorithm for Clustering Analysis of Prostate Cancer Patients:

Step 1 – Initialization:

Select a tolerance >0. Select an n-dimensional starting point $x^0 = (x^0_1,\ldots, x^0_n)$ and solve the following minimization problem:

$$\min f(x) = \sum_{i=1}^{r} \left\| x - a^i \right\| \quad s.t. \quad x \in \Re^n \tag{6}$$

Let an n-dimensional point x^{1*} be a solution to this problem and f^1 be the corresponding objective function value. Set $k=1$.

Step 2 – Computation of the next cluster:

Select an n-dimensional vector x^0, construct a new $2n$-dimensional starting point $x^{02}=(x^{1*},x^0)$ and solve the following optimization problem:

$$\min f^k(x) \quad s.t. \quad x \in \Re^n \tag{7}$$

where

$$f^k(x) = \sum_{i=1}^{m} \min\left\{ \left\| x^{1*} - a^i \right\|, \ldots, \left\| x^{k*} - a^i \right\|, \left\| x - a^i \right\| \right\}.$$

Step 3:

Let $x^{k+1,*}$ be a solution to the problem (7). Take $x^{k0}=(x^{1*},\ldots,x^{k*},x^{k+1,*})$ as a new starting point and solve the following minimization problem:

$$\min f^k(x) \quad s.t. \quad x \in \Re^{(k+1)\times n} \tag{8}$$

$$f^k(x) = \sum_{i=1}^{m} \min_{j=1,\ldots,k+1} \left\| x - a^i \right\|. \tag{9}$$

where

Step 4 – Termination Criterion:

Let $x^{k+1,*}$ be a solution to the problem (8) and $f^{k+1,*}$ be the corresponding value of the objective function. If

$$\frac{f^{k,*} - f^{k+1,*}}{f^{1,*}} < \varepsilon \tag{10}$$

then **Stop**, otherwise set $k=k+1$ and go to **Step 2**.

Both problems (7) and (8) are non-smooth optimization problem and the discrete gradient method developed by Bagirov [2] is used to address these problems.

4 Results and Discussion

Applying the optimization algorithm discussed in the previous section to the WBRC prostate cancer dataset with =0.01 using the patient's age, Gleason score, tumor stage, and PSA level at diagnosis as input parameters, ten clusters are produced as summarized in Table 4. Note that the "tumor stage" field was preprocessed by converting it to the scale of 1-8 where score 1 represents tumor stage 1a and score 8 represents tumor stage 4. The values reported represent the arithmetic mean (S), minimum and maximum value of input parameters $[m,M]$, and the coefficient of variation $K\%$ for all input parameters for every cluster obtained.

Note that clusters 2 and 3 that contain 2 and 6 records respectively are formed due to unusally large readings of the PSA and represent statistical outliers of some kind, but due to the nature of the problem domain, these clusters can hardly be excluded or ignored as extreme values of some parameters (such as very high PSA readings in relatively young patients) may suggest very intensive treatment options for some patients.

At the next stage of the study, the corresponding bFFS rates were calculated for every cluster as presented in Table 5 (note that the clusters are sorted in decreasing order of bFFS readings).

In order to comply with the arrangements on risk assessment adopted in WBRC, the percentage levels of 5yr bFFS considered as boundaries between risk groups were kept the same. According to these arrangements, clusters 10 and 5 represent low risk group, clusters 6, 3, 4, and 1 clearly include high risk patients, while the remaining clusters include the patients with intermediate risk. Note that the bFFS levels separation between clusters is much stronger pronounced than in the case of Table 3. This allows, for example, to closely investigate the patients from cluster 7 as they are much more likely to be at the riskier end of the intermediate group than other patients.

To use the same baseline for comparison, 10 clusters produced by the optimization algorithm should now be aggregated into three groups as summarized in Table 6.

Observe that using optimization based clustering approach both conditions 1 and 2 for an intuitively good classification discussed in section 2 are satisfied. While the percentage of biochemical failure free survival is similar for both initial WBRC approach and clustering approach, the resolution capacity of clustering approach is definitely much higher. In particular, 26% of patients are classified as low risk as opposed to 8% in the initial WBRC classification – these findings are of extreme importance for clinical judgement as for a reasonably large number of elderly people they may mean lower dosage of radiation therapy and/or the absence of other kinds of therapies with significant side effects. Note also that the improvement in the low risk group is obtained without any reduction effect on the high risk group and only by the means of reducing the intermediate risk group.

For many clinicians however, it is still difficult to trust the decisions of an automated algorithm, and the importance of rules to explain the decisions is critical to the acceptance of the data mining technology. Certainly, automated rule generation methods such as classification and regression trees (CART) are available to find rules describing different (pre-defined) subsets of the data. When the data sample size is limited though, such approaches tend to find very accurate rules that apply to only a small number of patients, who are frequently outliers. Clinicians are unlikely to trust rules generated under these conditions, particularly when existing classification rules are based on expert clinical knowledge.

This study demonstrates that data mining techniques can play an important role in rule refinement, even if the sample size is limited. Instead of using directed knowledge discovery (a.k.a supervised learning methods) to generate rules or models to assign prostate cancer patients into risk-homogeneous patient groups, we propose to use undirected knowledge discovery (a.k.a. unsupervised learning methods) to group the patients, explore the existing rules and identify possible inconsistencies and refinements.

5. Summary and Conclusions

In this paper a non-smooth non-convex optimization-based algorithm for solving cluster analysis problem is applied to mine the prostate cancer patients database consisting of 258 records for the purposes of generating new knowledge about patient risk groupings. It is demonstrated that the proposed approach can support clinical decision-making by improving the accuracy of risk assessment and it can, therefore, be seen as an evidence-based predictive tool

with high-knowledge generation capabilities.

As the proposed optimization algorithm calculates clusters step by step and the form of the objective function allows the user to significantly reduce the number of instances in a dataset, it can be effectively utilized for clustering in large-scale data sets. Conducting a similar study on a much larger prostate cancer database would therefore allow to overcome some of the limitations of this study that are due to the modest size of the WBRC database.

References

1. ASTRO: Consensus statement: guidelines for PSA following the radiation therapy. American Society for Therapeutic Radiology and Oncology Consensus Panel. International Journal of Radiation Oncology, Biology, and Physics 37(5) (1997) 1035-1041
2. Bagirov, A.M.: Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices, In: Eberhard, A., Hill, R., Ralph, D. and Glover, B.M. (eds.): Progress in Optimization: Contribution from Australasia, Kluwer Academic Publishers (1999) 147-175.
3. Bagirov, A.M. and Churilov, L.: An optimization-based approach to patient grouping for acute healthcare in Australia. The proceedings of the ICCS 2003 – The 3rd International Conference on Computational Science, Melbourne-St.Petersburg, Australia-Russian Federation (2003)
4. Bagirov, A.M, Rubinov, A.M., Yearwood, J.: A heuristic algorithm for feature selection based on optimization techniques. In: Sarker, R., Abbas, H. and Newton, C.S. (eds.): Heuristic and Optimization for Knowledge Discovery, Idea Publishing Group, Hershey (2002) 13-26
5. Bellazzi, R., Azzini, I., Toffolo, G., Bacchetti, S., Lise, M.: Mining data from a knowledge management perspective: an application to outcome prediction in patients with resectable hepatocellular carcinoma, In: Quaglini, S., Barahona, P., Andreassen, S. (eds): Artificial Intelligence in Medicine. 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001. Proceedings (Lecture Notes in Artificial Intelligence Vol.2101). Springer-Verlag. (2001), 40-49
6. Jain, A.K., Murty, M.N. and Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31(3) (1999) 264-323
7. Keeney, R.: Common mistakes in making value trade-offs. Operations Research, 50(6), (2002) 935-945
8. Land, W.H.Jr., Masters, T., Lo, J.Y., McKee, D.W., Anderson, F.R.: New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data, In: Embrechts, M.J., VanLandingham, H.F., Ovaska, S.J. (eds): SMCia/01 - Proceedings of the 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications. IEEE (2001) 47-52
9. Mangasarian, O.L.: Mathematical programming in data mining. Data Mining and Knowledge Discovery 1 (1997) 183-201
10. Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4) (1995) 570-577
11. Setiono, R.: Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in Medicine, vol.18, no.3, March, (2000) 205-219.
12. Walter, D. and Mohan, C.K.: ClaDia: a fuzzy classifier system for disease diagnosis. Proceedings of the 2000 Congress on Evolutionary Computation. IEEE. Part vol.2 (2000) 1429-1435
13. Zhang Z. and Zhang, H.: Development of a neural network derived index for early detection of prostate cancer. Proceedings of IJCNN'99 - International Joint Conference on Neural Networks. IEEE. Part vol.5 (1999) 3636-3641
14. Zupan, B., Demsar, J., Kattan, M.W., Beck, J.R., Bratko, I.: Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial Intelligence in Medicine, vol.20, no.1, Aug, (2000) 59-75

Table 1. William Buckland Radiotherapy Center (WBRC) risk group classification

Risk Group	Gleason Score	Tumor stage	PSA	5 yr bFFS
Low	2-6	T1c – 2a	0 - 10	**76.2%**
Intermediate	2-6	T1c – 2a	> 10 – 20	**53.4%**
	2-6	T2b – 3a	0 – 20	
	7	Up to T3a	0 – 20	
High	2-7	Up to T3a	> 20	**37.8%**
	2-7	T3b – 4	Any	
	8-10	Any	Any	

Table 2. WBRC risk group classification with relative sizes of the groups

	Low Risk	Intermediate Risk	High Risk
5yr bFFS %	76.2	53.4	37.8
% of records	8.2	51.2	38.3

Table 3. Rule-based 7 groups classifiaction of WBRC data

Risk groups	Low	Int1	Int2	Int3	High1	High2	High3
5 yr bFFS %	76	68	65	41	31	56	40
% of records	8.2	10.9	14.5	25.8	20.3	6.3	11.7

Table 4. Ten clusters produced by the Optimization algorithm

Cl	Size	Tumor stage			Gleason score			PSA			Age		
		S	***mM***	***K %***	***S***	***mM***	***K %***	***S***	***mM***	***K %***	***S***	***mM***	***K %***
1	48	7.2	6 8	6	7	5 9	12	9.3	2.41 5	38	68	51 81	10
2	2	7.0	7 7	0	7	7 7	0	186	149 243	28	70	66 73	7
3	6	6.7	4 8	20	6.7	4 8	22	77	61 99	20	69	58 81	11
4	19	6.2	3 8	25	6.9	6 9	12	43	35 54	14	69	60 76	6
5	41	3.4	1 5	29	5.9	2 10	23	4.8	0.6 9	49	65	44 78	11
6	29	4.1	2 5	19	6.4	4 9	20	25	20 36	16	69	52 79	8
7	60	4.6	4 5	10	6.8	5 9	14	12	5 19	25	70	57 83	9
8	18	7.3	7 8	6	6.7	4 8	14	24	17 34	21	68	56 75	7
9	9	6.2	4 8	24	2.9	2 4	29	11	7 19	34	69	55 78	10
10	26	3.2	1 4	21	5.7	3 7	17	15	10 20	18	67	54 76	8

Table 5. Ten clusters produced by the Optimization algorithm with corresponding bFFS values

Cluster	**10**	**5**	**9**	**8**	**2**	**7**	**1**	**6**	**3**	**4**
% of records	10	16	3	7	1	23	19	11	2	7
5 yr bFFS %	73	71	67	61	50	47	44	38	17	16

Table 6. Comparative analysis of clustering with optimization approach and the baseline case (WBRC rules)

	% 5 yr bFFS		**% records**	
	WBRC	**Clustering**	**WBRC**	**Clustering**
Low	76.2	71.6	8.2	26.0
Intermediate	53.4	53.1	51.2	34.0
High	37.8	35.7	38.3	39.0